Quantifying uncertainty in predictions of groundwater levels using formal likelihood methods

Ben Marchant[a*], Jonathan Mackay[a] and John Bloomfield[b]

[1]British Geological Survey, Environmental Science Centre, Keyworth, Nottingham, NG12 5GG, UK

[2]British Geological Survey, Maclean Building, Crowmarsh Gifford, Wallingford, Oxfordshire, OX10 8BB, UK

[*]Corresponding author: benmarch@bgs.ac.uk

Abstract

Informal and formal likelihood methods can be used to quantify uncertainty in modelled predictions of groundwater levels (GWLs). Informal methods use a relatively subjective criterion to identify sets of plausible or behavioural parameters of the GWL models. In contrast, formal methods specify a statistical model for the residuals or errors of the GWL model. The formal uncertainty estimates are only reliable when the assumptions of the statistical model are appropriate.

We apply the formal approach to historical reconstructions of GWL hydrographs from four UK boreholes. We test whether a model which assumes Gaussian and independent errors is sufficient to represent the residuals or whether a model which includes temporal autocorrelation and a general non-Gaussian distribution is required. Groundwater level hydrographs are often observed at irregular time intervals so we use geostatistical methods to quantify the temporal autocorrelation rather than more standard time series methods such as autoregressive models.

According to the Akaike Information Criterion, the more general statistical model better represents the residuals of the GWL model. However, no substantial difference between the accuracy of the GWL predictions and the estimates of their uncertainty is observed when the two statistical models are compared. When the general model is applied, significant temporal correlation over periods ranging from 3 to 20 months is evident for the different boreholes. When the GWL model parameters are sampled using a Markov Chain Monte Carlo approach the distributions based on the general statistical model differ from those of the Gaussian model, particularly for the boreholes with the most autocorrelation. These results suggest that the independent Gaussian model of residuals is sufficient to estimate the uncertainty of a GWL prediction on a single date. However, if realistically autocorrelated simulations of GWL hydrographs for multiple dates are required or if the distributions of the GWL model parameters are of interest, then the more general statistical model should be used.

Keywords

Groundwater, likelihood, mixed models, formal, MCMC

1. Introduction

Groundwater level (GWL) hydrographs from boreholes provide valuable information about the return periods and severity of past drought and flood events. There is often a need to use deterministic models to extend a hydrograph record either to reconstruct GWLs prior to the drilling of the borehole (Mackay et al., 2014), to interpolate GWLs on dates when they were not observed (Sun et al., 2009) or to forecast future GWLs (Daliakopoulos et al., 2004). Reconstructed hydrographs might assist scientists in understanding the influence of long-term anthropogenic processes such as abstraction or climate change on the variation of GWLs (Shepley and Soley, 2012) and more specifically on the severity of extreme events (Kidmose et al., 2013). Interpolations of incomplete hydrograph records are required to compute standardised indices of GWLs (e.g. Bloomfield and Marchant, 2013) that can place current GWLs in a historical context. Forecasts of GWLs might warn land managers and policy makers of potential extreme events so that remediation efforts are focussed appropriately (Jackson et al., 2013).

In all of these contexts, it is vital that the uncertainties of the modelled GWLs are quantified so that land managers and scientists interpreting predictions can determine which features reflect statistically significant variation in groundwater processes rather than model errors (Jackson et al., 2015). Uncertainties can arise because of errors in inputs to the groundwater models such as rainfall amounts, errors in the model structure, measurement errors and errors in the estimated parameters of the groundwater model. Schoups and Vrugt (2010) describe two approaches for estimating the parameter uncertainty of hydrological models. Both approaches lead to an ensemble of plausible parameters rather than a single optimal set. With formal likelihood methods, a statistical model for the residuals is specified and used to derive a likelihood function to quantify the probability that the observed data would have arisen from the hydrological model with a particular parameter set. The likelihood function is then used to assess which parameter sets are plausible. The framework which combines a deterministic model with a statistical model of the residuals is referred to as a mixed model (MM; Dobson, 1990). The predictions from the deterministic model constitute the fixed effects and the predictions of the residual model are the random effects. The random effects will include contributions resulting from input errors, measurement errors and model structural errors. A MM can be used to predict the entire probability density function (pdf) for the property of interest on a target date when it was not observed (Lessels and Bishop, 2013). This pdf is conditioned on the available observations and accounts for the correlation between the random effects on the target and observation dates.

Beven et al. (2008) note that the formal likelihood approach relies on the assumptions of the statistical model and these assumptions might be inappropriate. For example, a simple model for the random effects might specify that they are independent and realized from a Gaussian distribution with zero mean and constant (i.e. stationary) variance. However, autocorrelated, non-Gaussian errors with non-stationary variance often occur (Kuczera, 1983). Therefore, Beven et al. (2008) advocate informal likelihood methods such as generalized likelihood uncertainty estimation (GLUE; Beven and Freer, 2001). In the GLUE approach the likelihood function for the errors is not linked to a specific error model. Instead, metrics such as the proportion of the variance of the observations that is explained by the hydrological model, are used to assess the plausibility of a particularly set of parameters. Thus the GLUE approach is free from assumptions but a somewhat subjective choice of likelihood function is required. The identified set of plausible parameter values can be used to generate multiple modelled reconstructions of the property of interest. The between-reconstruction variability corresponds to the contribution of parameter uncertainty to the total uncertainty. However, there is no model to predict the other contributions such as input uncertainty and model structural errors.

Schoups and Vrugt (2010) respond to the concerns of Beven et al. (2008) by generalizing their random effects models to accommodate non-Gaussian variation, temporal correlation and non-stationary variances. They assume that their random effects are realized from a skew exponential power (SEP) distribution which specifies the skewness and kurtosis of the residuals independently. The SEP distribution permits more general non-Gaussian variation than a transformation of the observed data (e.g. Box and Cox, 1964). The variance of the residuals is permitted to vary according to streamflow and the temporal correlation is represented by autoregressive time series models (Chandler and Scott, 2011). Schoups and Vrugt (2010) demonstrated their approach on rainfall runoff models in both humid and arid basins. They used Bayesian uncertainty methods to sample plausible sets of parameters for both the fixed and random effects models. For both basins, the observed flow data had a very heavy tail which resulted from the large and rapid response of the flow to large storm events. Schoups and Vrugt (2010) found that their generalized non-Gaussian model led to larger likelihoods than a model which assumed independent and Gaussian random effects. Hence the assumption of independent and Gaussian residuals was not appropriate. The generalized random effects model did not improve the model predictions. In fact, the mean squared errors were smaller for the independent Gaussian model because the generalized model led to more emphasis being placed on accurately estimating the low rather than large flows. The generalized model did however lead to large improvements in the estimates of the uncertainty of the model predictions. Also, the inappropriate Gaussian model led to quite different distributions of plausible parameters than were realized from the generalized model.

The formal and informal likelihood approaches are also applicable to groundwater models. Jackson et al. (2016) use the GLUE methodology to assess the uncertainty of model reconstructions of groundwater levels at six boreholes in the UK. The reconstructions are generated using the AquiMod conceptual model (Mackay et al., 2014) and the Nash Sutcliffe Efficiency (NSE) score (Nash and Sutcliffe, 1970) is used to decide which sets of parameters are plausible. Von Asmuth and Bierkens (2005), Mirzavand and Ghazavi (2015) and Peterson and Western (2014) all specify a formal statistical model for the residuals from their models of GWLs. These statistical models are based on a Gaussian distribution but they account for temporal autocorrelation amongst the residuals.

In this paper we quantify the uncertainty of AquiMod reconstructions for GWLs in four English boreholes using a formal likelihood approach similar to Schoups and Vrugt (2010) and we discuss the relative suitability of the formal and informal approaches for quantifying the uncertainty of UK groundwater models. We also explore whether the assumption of independent and Gaussian residuals is suitable for a formal model of GWLs in this context or whether a more general model is required. Groundwater hydrographs tend to be less heavy tailed than river hydrographs since, in effect, the hydrogeological system acts as a filter which temporally smooths the effects of intense storm events.

One modification to the approach Schoups and Vrugt (2010) is necessary. The autoregressive models which they use to represent temporal autocorrelation amongst the residuals are well-suited to data observed at regular intervals but they cannot be applied to irregularly sampled time series. Until the relatively recent installation of automated telemetry in some groundwater boreholes, GWLs tended to be recorded at irregular time intervals (Environment Agency, 2014) according to factors such as the availability of staff to visit the borehole and conduct a dip-test. Von Asmuth and Bierkens (2005) recognised this problem and suggested a continuous approximation to the autoregressive model. This approach was adopted by Peterson and Western (2014). In the more general hydrological context, Chandler and Scott (2011) recommend that the temporal correlation amongst irregularly sampled water levels is represented by a variogram. Variograms are more commonly associated with spatial analyses (Webster and Oliver, 2007) and they describe how the expected squared difference between a pair of observations varies according to the lag between the observation sites or times. Chandler and Scott (2013) use the method of moments to estimate their variograms but this estimator is not immediately compatible with formal likelihood functions. Therefore, we consider model-based variogram estimators (Diggle and Ribeiro, 2007) that use formal likelihood functions. Lessels and Bishop (2013) used a linear MM with an exponential variogram to represent irregularly measured water quality parameters from two catchments in southeast Australia. We represent the variograms by the flexible four parameter Matérn function which generalises many commonly used variogram functions such as the exponential model (Marchant and Lark, 2007).

2. A mixed model for groundwater levels

We represent the GWLs observed at times $t = t_1, t_{2,} \ldots, t_n$ by a MM:

$$z(t_i) = m(t_i|\boldsymbol{\beta}_\mathbf{f}) + r(t_i|\boldsymbol{\beta}_\mathbf{r}), \tag{1}$$

where $m(t_i|\boldsymbol{\beta}_\mathbf{f})$ is the deterministic model prediction or fixed effect at time t_i and $r(t_i|\boldsymbol{\beta}_\mathbf{r})$ is the random effect or residual at time t_i. The $\boldsymbol{\beta}_\mathbf{f}$ and $\boldsymbol{\beta}_\mathbf{r}$ are the calibrated parameters for the fixed and random effects respectively. For brevity, we henceforth denote $z(t_i)$, $m(t_i|\boldsymbol{\beta}_\mathbf{f})$ and $r(t_i|\boldsymbol{\beta}_\mathbf{r})$ by z_i, m_i and r_i respectively.

2.1 The fixed effects model

Mackay et al. (2014) reviewed the types of models used to simulate GWLs. They distinguished 'black box' methodologies such as statistical transfer functions (e.g. Jakeman et al., 2006) from process-driven models based on simplifications of physical laws of fluid dynamics (e.g. Shepley and Soley, 2012). They noted that the process-driven models can relate the variation of GWLs to hydrogeological properties but that the calibration of these models is complex and requires more data than are typically available. In contrast, black box methodologies are more easily calibrated but they provide little insight into the controls on GWLs. Mackay et al. (2014) proposed AquiMod, a conceptual lumped parameter model, as a compromise approach. Rather than starting from basic physical principles, such conceptual models contain simpler representations of the components of the hydrogeological system. We use AquiMod as the fixed effects of our MM.

AquiMod includes simple conceptual representations of soil drainage, the transfer of water through the unsaturated zone and groundwater flow. The soil zone is represented as a bucket which receives water from rainfall and releases water through evapotranspiration and drainage into the unsaturated zone. Measured rainfall and potential evapotranspiration amounts are inputs to the model. Model parameters control the nonlinear relationship between the rate of evapotranspiration and the soil moisture and the rate at which water drains from the soil. A parametric transfer function is used to represent the rate of vertical flow through the unsaturated zone into the saturated zone which is represented by a layered rectangular block of aquifer through which water flows horizontally. Water discharges from a layer via an outlet at its base. The discharge rate is dependent on the layer permeability and the hydraulic gradient. The former is controlled using a hydraulic conductivity parameter and the latter is controlled by outlet elevation and aquifer length parameters. The number of layers is selected for each borehole to lead to the best possible match between observed and modelled GWLs.

This model is easier to calibrate than a more complex process-based model and more easily interpretable than a black-box model. An implementation of AquiMod with three layers in the saturated zone has a total of 16 parameters but eight of these can be estimated from available information about the catchment. The remaining eight parameter values constitute our fixed effects parameter vector $\boldsymbol{\beta}_{\mathbf{f}}$. These parameters are Z_r (mm) the maximum rooting depth of vegetation, p

the water depletion factor of vegetation, λ the scale parameter of a Weibull function which describes the rate of recharge, k the shape parameter for the Weibull function, S (%) the aquifer storage coefficient and k_i $i = 1,2,3$ (m d^{-1}) the hydraulic conductivity for each layer of the saturated zone. Mackay et al. (2014) estimated these parameters by finding the values which led to the largest Nash-Sutcliffe Efficiency (NSE) score. The NSE score is measure of the proportion of variance which has been explained by the model and is defined as:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}\{z_i - m_i\}^2}{\sum_{i=1}^{n}\{z_i - \bar{z}\}^2}, \tag{2}$$

where $\bar{z}$ is the mean of the observed values. Jackson et al. (2016) set a threshold of 0.5 on the NSE when deciding which parameter vectors were plausible.

2.2 The random effects model

We assume that the random effects, r_i, are a realization of a multivariate random function with zero mean. A simple random effects model would assume that the random function is Gaussian with fixed variance and that the realizations from this function are independent. However, this model might not be sufficiently flexible to represent observed GWLs (Bloomfield and Marchant, 2013). A more general random function can be described in terms of its marginal distributions and its dependence structure or copula (e.g. Bárdossy and Li, 2008). A marginal distribution is the pdf for a random function, in our case the random effects, at a single time. It does not take any account of the random effects at other times. In this paper, we assume that the residuals at each time are realized from the same marginal distribution with density $f(r)$ and cumulative distribution function (cdf) $F(r)$. Therefore, if an appropriate marginal distribution is specified, the set of $u_i = F(r_i)$ quantile values should be a sample from a uniform distribution bounded by zero and one. The copula describes the correlation between the u_i. If we assume a Gaussian copula and denote the cdf of a standardised Gaussian distribution by $\Phi_{0,1}$ then $\mathbf{a} = (a_1, \ldots, a_n)$, where $a_i = \Phi_{0,1}^{-1}(u_i)$, is a realization of a multivariate Gaussian distribution where each marginal has zero mean and unit variance and the a_i are linearly correlated with correlation matrix $\mathbf{C}$.

The log-likelihood of a multivariate random function with marginal distribution $f(r)$, Gaussian copula and correlation matrix $\mathbf{C}$ can be written (Kazianka and Pilz, 2010; Marchant et al., 2011)

$$l = -\tfrac{1}{2}\log|\mathbf{C}| + \tfrac{1}{2}\mathbf{a}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{C}^{-1})\mathbf{a} - \textstyle\sum_{i=1}^{n}\log[f(r_i)], \tag{3}$$

where $\mathbf{I}_n$ is the identify matrix of length n.

We relax the standard assumption of independent Gaussian residuals by calibrating random effects models with more general marginal distributions and correlation matrices. Our choice of marginal distribution is the asymmetric exponential power (AEP) distribution (Figure 1). This has density (Zhu and Zinde-Walsh, 2009):

$$f(x) = \begin{cases} \left(\dfrac{\alpha}{\alpha^*}\right)\dfrac{1}{\sigma}K_{\mathrm{EP}}(p_1)\exp\left(-\dfrac{1}{p_1}\left|\dfrac{x-\mu}{2\alpha^*\sigma}\right|^{p_1}\right) & \text{if } x \le \mu \\ \left(\dfrac{1-\alpha}{1-\alpha^*}\right)\dfrac{1}{\sigma}K_{\mathrm{EP}}(p_2)\exp\left(-\dfrac{1}{p_2}\left|\dfrac{x-\mu}{2(1-\alpha^*)\sigma}\right|^{p_2}\right) & \text{if } x > \mu. \end{cases} \tag{4}$$

where μ is the location parameter, $\sigma > 0$ is the scale parameter, $\alpha \in (0,1)$ is the skewness parameter, $p_1 > 0$ and $p_2 > 0$ are the left and right tail parameters, $K_{\mathrm{EP}}(p) = 1/\left[2p^{1/p}\Gamma(1+1/p)\right]$ is the normalizing constant, Γ is the Gamma function and $\alpha^* = \alpha K_{\mathrm{EP}}(p_1)/[\alpha K_{\mathrm{EP}}(p_1) + (1-\alpha)K_{\mathrm{EP}}(p_2)]$. Figure 1 illustrates how the skewness and the decay of the left and right tails of the pdf are controlled by α, p_1 and p_2. When $p_1 = p_2$ the AEP distribution reduces to an alternative parameterisation of the SEP distribution used by Schoups & Vrugt (2010). Zhu and Zinde-Walsh also derive expressions for the AEP cdf in terms of the Gamma cdf $\mathrm{G}(x,\gamma)$:

$$F(x) = \begin{cases} \alpha\left[1 - G\left(\dfrac{1}{p_1}\left(\left|\dfrac{x-\mu}{2\alpha^*\sigma}\right|\right)^{p_1}, \dfrac{1}{p_1}\right)\right] \text{ if } x \le \mu \\ \alpha + (1-\alpha)G\left(\dfrac{1}{p_2}\left(\left|\dfrac{x-\mu}{2(1-\alpha^*)\sigma}\right|\right)^{p_1}, \dfrac{1}{p_2}\right) \text{ if } x > \mu \end{cases}, \tag{5}$$

the quantile function of the AEP:

$$F^{-1}(v) = \begin{cases} \mu - 2\sigma\alpha^* \left[p_1 G^{-1}\left(1 - \frac{v}{\alpha}, \frac{1}{p_1}\right)\right]^{\frac{1}{p_1}} \text{ if } v \leq \mu \\ \mu + 2\sigma(1-\alpha^*) \left[p_2 G^{-1}\left(1 - \frac{1-v}{1-\alpha}, \frac{1}{p_2}\right)\right]^{\frac{1}{p_2}} \text{ if } v > \mu \end{cases}, \quad (6)$$

and demonstrate that the expectation of an AEP distributed random variable X is:

$$\mathrm{E}(x) = \mu + \frac{\sigma}{B}\left[(1-\alpha)^2 \frac{p_2\Gamma(2/p_2)}{\Gamma^2(1/p_2)} - \alpha^2 \frac{p_1\Gamma(2/p_1)}{\Gamma^2(1/p_1)}\right], \quad (7),$$

where $B = \alpha K_{\mathrm{EP}}(p_1) + (1-\alpha)K_{\mathrm{EP}}(p_2)$.

Schoups and Vrugt (2010) use autoregressive models to determine the correlation between random effects. However, this approach cannot be used when the GWLs are observed at irregular time intervals. Instead, we use a model-based geostatistical approach (Diggle and Ribeiro, 2007) and calculate the entries of $\mathbf{C}$ using an authorized parametric function which ensures that $\mathbf{C}$ is positive definite. Many such authorized functions exist (Webster and Oliver, 2007) including the exponential model which is a continuous equivalent to an autoregressive model of order 1. We choose the more general Matérn function:

$$\mathbf{C}_{i,j} = \begin{cases} 1 & \text{if } |t_i - t_j| = 0 \\ \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{|t_i - t_j|}{\rho}\right)^{\nu} K_\nu\left(\frac{|t_i - t_j|}{\rho}\right) & \text{if } |t_i - t_j| > 0 \end{cases} \quad (8)$$

to express entry i, j of the covariance as a function of $|t_i - t_j|$, the time separating the two observations. The Matérn function has two parameters, namely the distance parameter ρ and the smoothness parameter ν. The smoothness parameter controls the rate of decay of the function for small lags (Figure 2). When $\nu = 0.5$ the Matérn function is equal to the exponential function. If we select μ to ensure that $\mathrm{E}(r) = 0$ then our general model of the residuals has six parameters to be calibrated i.e. $\boldsymbol{\beta}_{\mathbf{r}} = (\sigma, \alpha, p_1, p_2, \rho, \nu)$.

2.3 Calibration of the mixed model

Our general MM has a total of 14 parameters which must be estimated or calibrated on observed GWLs. The maximum likelihood (ML) estimator uses a numerical optimization algorithm to find the values of these parameters that maximizes the log-likelihood function (Eqn. 3) for the calibration data. It is also possible to compare different model structures by comparing their likelihoods. For example, one might wish to consider whether an AquiMod fixed effects model with a three layer saturated zone is a significantly better fit to the data than a model with only two layers. Alternatively, one might consider whether a random effects model with a Matérn correlation function is superior to one that uses an exponential model. This can be achieved by fitting each model by ML and then comparing their Akaike Information Criterion (AIC; Akaike, 1973):

$$\mathrm{AIC} = 2k - 2l, \tag{14}$$

for the maximized log-likelihood l from Equation 3. Here, k is the total number of parameters in the model. If too many parameters are included in a model it might be over-fitted. This means that the model matches the intricacies of the calibration data very closely but is less suitable for representing independent validation data. The model with the lowest AIC is thought to be the best compromise between complexity and quality of fit to the data (Webster and Oliver, 2007). Alternative information criteria such as the Bayesian Information Criterion (BIC, Marshall et al., 2005) do exist. However, the formula for the BIC includes the number of observations. When the observations are highly correlated some adjustment of this term will be required.

The ML estimator is a frequentist method that looks for the single set fixed parameter values that generated the observed data (Minasny et al., 2011). In reality, such a set of parameters rarely exist. Deterministic models tend to approximate the complexities of environmental systems. Even if an optimal deterministic model existed, it is highly unlikely that sufficient calibration data would be available to uniquely identify the parameters of this model. Indeed, many deterministic models include state variables that are unmeasurable. Therefore a number of models are likely to be suitable to represent the environmental system (Beven and Binley, 1992).

In a Bayesian analysis the model parameters are treated as probabilistic variables. Our knowledge of the parameter values prior to collecting any data is expressed as a prior distribution. Then the observations are used to update these priors and to form a posterior distribution of the GWLs which combines our prior knowledge with the information that could be inferred from the observations. The posterior distribution can be sampled using a Markov Chain Monte Carlo (MCMC) approach (Diggle and Ribeiro, 2007). This is an iterative method which moves between behavioural or plausible parameter vectors according to the corresponding values of the log-likelihood function. The parameter vector is randomly perturbed and the Metropolis-Hastings algorithm (Hastings, 1970) is used to decide if the parameter set is behavioural. The MCMC approach is computationally demanding and the perturbations of the parameter vector must be carefully controlled to ensure that the sample is representative of the behavioural parameter set. Until recently, these challenges might have prevented the use of the approach to estimate all of the parameters of a deterministic GWL model. However, Vrugt et al. (2008) have developed the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm which permits efficient MCMC sampling in high dimensional parameter spaces and automatically selects effective perturbations of the parameter vector. Minasny et al. (2011) demonstrated how this algorithm could be applied in conjunction with geostatistical models and Minasny et al. (2013) described how it could be extended to MMs that included nonlinear fixed effects models. The MCMC sample can also be used to assess whether the parameters of the fixed effects model are identifiable. This concept is formally defined by Renard et al. (2010). A parameter is non-identifiable if the observed data do not provide any information about that parameter. In this case the posterior distribution of the parameter is no more certain than the prior distribution.

2.4 Prediction using the mixed model

The calibrated mixed model can be used to make a prediction of the pdf of the residuals or to generate simulations of the residuals at a set of times where the GWL was not observed. These predictions and simulations are conditional on the observations that are available. In our discussion above we demonstrated that $\mathbf{a} = (a_1, \dots, a_n)$, where $a_i = \Phi_{0,1}^{-1}(u_i)$ is a realization of a multivariate Gaussian random function. The kriging predictor (Webster & Oliver, 2007) can be used to predict a at a set of q target times $\mathbf{T} = (T_1, \dots, T_q)$ when it was not observed. We denote these predictions by $\mathbf{a}_\mathrm{T}$. The length q vector of expectations $\mathbf{e}_\mathrm{T}$ and the $q \times q$ covariance matrix $\mathbf{v}_\mathrm{T}$ of $\mathbf{a}_\mathrm{T}$ are:

$$\mathbf{e}_\mathrm{T} = \mathbf{C}_\mathrm{TO}\mathbf{C}^{-1}\mathbf{a}, \tag{9}$$

$$\mathbf{v}_\mathrm{T} = \mathbf{C}_\mathrm{TT} - \mathbf{C}_\mathrm{TO}\mathbf{C}^{-1}\mathbf{C}_\mathrm{TO}^\mathrm{T}, \tag{10}$$

where $\mathbf{C}_\mathrm{TO}$ is the $q \times n$ matrix of correlations between the residuals on the target times and the observed residuals and $\mathbf{C}_\mathrm{TT}$ is the correlation matrix for the residuals at the target times. The LU method (Webster & Oliver, 2007) can be used to generate simulations of $\mathbf{a}_\mathrm{T}$. These can be transformed to simulations of the residuals:

$$\mathbf{r}_\mathrm{T} = F^{-1}\left[\Phi_{0,1}(\mathbf{a}_\mathrm{T})\right]. \tag{11}$$

The correlation between the elements of each of the realizations will be consistent with the calibrated model. Alternatively it is possible to predict the pdf of the residual for a single target time conditional on the observed residuals $\mathbf{r}$ by calculating:

$$f(r^*|\mathbf{r}, \boldsymbol{\beta}_\mathbf{r}) = \frac{f(r^*) \times \phi_{e_T,\sqrt{v_T}}(a^*)}{\phi_{0,1}(a^*)}, \tag{12}$$

for the range of plausible values of r^*. Here, $a^* = \Phi_{0,1}^{-1}[F(r^*)]$, $\phi_{e,b}$ is the pdf of a Gaussian distribution with mean e and standard deviation b and e_T and v_T are the expectation and variance of the kriged prediction on this single date. Note that if the residuals on the target date are independent of the conditioning observed residuals or if no conditioning observations are included in the prediction, then $e_\mathrm{T} = 0, v_\mathrm{T} = 1$ and Eqn. 12 reduces to $f(r^*|\mathbf{r}, \boldsymbol{\beta}_\mathbf{r}) = f(r^*)$.

If the MM has been calibrated by the MCMC approach then an ensemble of independent realisations of the parameter vector will have been sampled. The pdf of the residual conditional on the observed GWLs is then equal to Eqn. 12 averaged across the parameter vectors. This pdf accounts for the uncertainty in estimating the fixed and random effects parameters and the residual errors of the fixed effects model.

2.5 Validation of the mixed model

The NSE score (Eqn. 2) is commonly used as a criterion to validate hydrological models. However, Thyer et al. (2009) note that the NSE score is only a measure of the accuracy of the predictions and it cannot be used to confirm that the assumed distribution of the random effects is consistent with the observed data. Therefore, Thyer et al. (2009) recommend the use of the predictive QQ plot. If the calibrated MM is used to predict the ith observed GWL then the p-value of the observed value is equal to:

$$p_i = \Phi_{0,1}\left(\frac{\tilde{a}_i - e_i}{\sqrt{v_i}}\right), \tag{13}$$

where $\tilde{a}_i$ is the observed value of a at time i (i.e. $\tilde{a}_i = \Phi_{0,1}^{-1}[F(\tilde{r}_i)]$), and e_i, v_i are the expectation (Eqn. 9) and variance (Eqn. 10) of the kriged prediction of a_i. If the observed GWL is a realization of the MM then p_i is a realization of a uniform distribution on $[0,1]$. A QQ plot is constructed by calculating p_i for a large number of observations. The p_i are sorted and plotted against the theoretical p-values or the cdf of the uniform distribution (i.e. evenly spaced values between zero and one). If all of the points of the QQ plot lie on the 1:1 line, the MM predictive distribution agrees exactly with the observations. If all the points lie above (or alternatively, below) the 1:1 line then the GWLs are under (over) predicted. If the points lie below (above) the 1:1 line for small theoretical p-values and above (below) the 1:1 line for large theoretical p-values then the predictive uncertainty of the MM is under (over) estimated.

3. Methods

All of the computations were conducted using Matlab (Mathworks, 2014) and the Matlab implementation of the DREAM algorithm (Vrugt, 2016) was used to perform the MCMC analyses.

3.1 Borehole and Meteorological Data

We estimated MMs for the monthly records from four boreholes considered by MacKay et al. (2014). These boreholes are named Chilgrove House (540 observations), Hucklow South (440 observations), Lower Barn Cottage (368 observations) and Skirwith (326 observations) and they are set in chalk, limestone, lower greensand and sandstone respectively. Three of these boreholes were considered by Jackson et al. (2016) when they used GLUE to quantify uncertainty in a GWL model. MacKay et al. (2014) give full details about the characteristics and setting of the boreholes. The GWL records were extracted from the UK National Groundwater Archive (National Groundwater Level Archive, 2013). Monthly GWLs from the boreholes are shown in Figure 3. Strong seasonal patterns are evident in all of the hydrographs and the Lower Barn Cottage and Skirwith hydrographs are considerably smoother than the other two. We follow MacKay et al. (2014) and use the first half of these time series for calibration of our MMs and the second half for validation. There are 18 missing observations from Skirwith during the validation period.

The monthly precipitation data required as an input to AquiMod were extracted from the Centre for Ecology and Hydrology's CERF 1km gridded precipitation dataset which is derived from UK Meteorological Office data (Keller et al., 2005). The monthly potential evapotranspiration time series were extracted from the Meteorological Office Rainfall and Evaporation Calculation System (Field, 1983). These are based on a modified version of the Penman-Monteith equation (Monteith and Unsworth, 2008).

3.2 Model calibration and analyses

We calibrated a series of MMs with increasingly complex random effects using the ML estimator. The initial models assumed that the random effects were independent and realized from a Gaussian distribution with constant variance. This Gaussian model was then generalized to include temporal correlation described firstly by exponential and then by Matérn covariance functions. The independent, exponential and Matérn covariance models were then used in conjunction with the AEP distribution. The structure of the fixed effects models were identical to the AquiMod models used by MacKay et al. (2014). The Chilgrove House and Hucklow South models had three saturated zone structures whereas there were only two saturated zone structures for Lower Barn Cottage and Skirwith. The AIC was calculated for each calibrated MM. For each borehole, the MM with the lowest AIC was used to predict the GWLs during the validation period and to calculate the uncertainty of these predictions. Predictive QQ plots were calculated for the calibration observations (without conditioning data) and the validation observations.

For each borehole, the MMs were recalibrated using the DREAM MCMC approach. All of the AquiMod parameters were assumed to have uniform prior distributions. The bounds on these parameters were identical to the parameter ranges considered by MacKay et al. (2014). The MCMC was iterated 600 000 times. Initial runs of the algorithm indicated that around 15 000 iterations were required before the Gelman-Rubin convergence diagnostic (Vrugt et al., 2009) was consistently less than 1.2 indicating that the MCMC had converged to the plausible portion of the parameter space. There was evidence of some correlation between sampled parameter vectors separated by up to 100 iterations. We therefore conservatively discarded the first 100 000 parameter vectors and only selected every 500th on the remaining vectors to yield an ensemble of 1 000 parameter vectors which we treated as if they were independent samples of the parameter set. The validation procedures that had been applied to the ML estimates were then repeated for the ensembles of MCMC parameter estimates. We also used the MCMC ensembles to assess which of the AquiMod parameters were identifiable.

4. Results

4.1 Maximum likelihood estimation of the mixed models

Table 1 shows the AIC values for the ML estimates of the different MMs for each borehole. In each case, the inclusion of the AEP rather than Gaussian random effects and the inclusion of autocorrelated random effects led to a decrease in the AIC. In contrast the NSE scores (Table 2) were largely unchanged as MM complexity was increased. Indeed, in the case of Lower Barn Cottage there is a sharp decrease in NSE when the AEP random effects with exponential correlation function are generalised to a Matérn correlation function. For three of the four boreholes the lowest AIC was achieved with an exponential covariance function but at Lower Barn Cottage there was sufficient improvement in likelihood to justify the use of a Matérn function.

The predicted pdfs of the random effects based on these best fitting MMs varied in terms of the magnitude and direction of their skew and the rate of decay of each tail (Figure 4). However, the observed residuals from the fixed effects model were generally consistent with these predicted pdfs. The best fitting models also differ in terms of their autocorrelation functions (Figure 5). At Chilgrove House and Hucklow South temporal autocorrelation is only evident for time lags of less than 5 months whereas for Skirwith there is correlation for lags up to 20 months and for Lower Barn Cottage, observations separated by well over 20 months are autocorrelated.

The predictions of GWLs for the four sites during the validation period (Figure 6) followed the same pattern of peaks and troughs as the observed values and the observations were generally within the 90% confidence limits of the predictions. The predictive QQ plots for the calibration data at Chilgrove House, Hucklow South and Skirwith were all reasonably close to the 1:1 line for both the Gaussian independent random effects (Figure 7) and the best fitting general random effects model (Figure 8). However, the corresponding plots for Lower Barn Cottage were further from the 1:1 line, particularly for the more general model where the curve was consistently above the 1:1. This indicates that there is systematic under prediction of GWLs and is consistent with the relatively poor NSE score for this site. We suspected that there were too few observations to accurately estimate all of the components of the general MM for Lower Barn Cottage. Also, the observations that were available were highly correlated. Therefore, we re-calibrated the MM for this site using both the original calibration and validation observations and saw a marked improvement in the QQ plot (Figure 8e). At Chilgrove House, the QQ plot for the validation data was also close to the 1:1 line for both the Gaussian and general random effects models. There was some moderate under estimation of the uncertainty of the validation predictions using the general random effects model at Hucklow South and slightly more severe over-estimation of the uncertainty at Skirwith. The validation QQ plots for both the Gaussian and best fitting random effects models were both relatively poor for Lower Barn Cottage.

4.2 MCMC estimation of parameters

The ensembles drawn from the MCMC samplers indicate that the AquiMod parameters (Figure 9) and parameters of the AEP distribution (Figure 10) are generally identifiable for each borehole. The parameters are confined to a range that is less than that of the prior distribution. Note that the range on the x-axis in these plots is identical to the range of the prior uniform distribution of the parameter. However, the spread of the posterior realizations of k, the shape parameter of the Weibull distribution within AquiMod, is almost as wide as the prior distribution for all of the boreholes. Closer inspection of the MCMC ensembles revealed that this parameter is highly correlated to λ, the scale parameter of the Weibull distribution, and this relationship explains the identifiability issue. At Lower Barn Cottage and Skirwith the identifiability of many of the AquiMod parameters improves when the entire observation record, rather than half of it, is used for calibration.

The posterior ensembles from the Gaussian independent random effects model (red histograms in Figure 8) and the best fitting random effects model (grey histograms) are relatively similar for

Chilgrove House and Hucklow South. However, for Lower Barn Cottage and Skirwith, there are marked differences between the posterior distributions of the parameters.

The MCMC ensembles of random effects parameters were used to estimate the uncertainty of the correlation functions (Figure 5). These plots illustrate that there is significant temporal auto-correlation amongst the random effects at the p=0.05 level for more than 2 months at Chilgrove House, more than 5 months at Hucklow South and more than 20 months at Lower Barn Cottage and Skirwith. Figure 11 shows histograms of NSE values achieved by each parameter vector within the MCMC ensembles for each borehole. For Chilgrove House and Hucklow South these values are fairly tightly clustered around the maximum. For Lower Barn Cottage and Skirwith the NSE values are more variable indicating that the proportion of GWL variation explained by the fixed effects models varies between different parameter vectors within the ensemble. The NSE values for these two boreholes do become more clustered close to the maximum when the entire observation record is used to calibrate the model.

5. Discussion

5.1 Overview

We have demonstrated how MMs can be used to represent the temporal variation of GWLs at specific boreholes and to predict these GWLs on dates where they were not measured. These predictions can be used to reconstruct hydrographs for times prior to the drilling of the borehole, to fill in gaps in the hydrograph through interpolation and to simulate potential future characteristics of hydrographs under different climate scenarios. The MM framework is flexible in terms of the deterministic model that may be included in the fixed effects and the structure of the random effects. The MMs were tested using the same monthly GWL observations that had previously been modelled by Mackay et al. (2014) using informal methods. However, the correlation functions included in the random effects are fully compatible with the irregularly sampled hydrographs that are available for many sites in the UK (Environment Agency, 2014). The uncertainty of MM parameters can be accounted for by sampling them using a MCMC approach. This reveals that they are generally identifiable although some parameters cannot be uniquely defined if they are strongly related to each other (Renard et al., 2009).

5.2 Structure of the random effects

For all four boreholes the best fitting MM according to the AIC included temporal autocorrelation amongst the random effects which were realized from an AEP rather than a Gaussian distribution. This indicates that the residuals are inconsistent with the assumptions that they were independent and realized from a Gaussian distribution. In this respect our results agree with the findings of Schoups and Vrugt (2010) for rainfall runoff models. Also in common with Schoups and Vrugt (2010), we found that the accuracy of the GWL predictions was not substantially improved by including the more general random effects model. The NSE scores achieved in this study were a very slight improvement on those recorded by Mackay et al. (2014) for the same boreholes but we suspect that these improvements were wholly due to differences in the numerical methods used to minimize the objective function when estimating the parameters rather than a difference in the modelling approach.

Schoups and Vrugt (2010) found that the inclusion of their general random effects model did lead to substantial improvements in the estimates of the predictive uncertainty. This did not occur for the predictions from AquiMod. This difference could have arisen because the deviations of GWLs from the Gaussian distribution are far less severe than for stream flows where very heavy tails result from sharp responses to storm events. Indeed, the models of Schoups and Vrugt (2010) included a relationship between the error variance and the flow rate but when we experimented with such relationships for AquiMod (results not shown) the likelihood did not improve.

The QQ plots indicated that the predictive uncertainty of AquiMod was relatively poorly estimated by the autocorrelated AEP models for the two sites with substantial temporal correlation. We suspect this was because there were too few observations from these sites and those that were available were too strongly correlated to accurately estimate all of the parameters of the MM. The likelihood function for an auto-correlated variable is known to be particularly sensitive to the correlation between close pairs of observations (Stein, 1999). It appears that when insufficient data are available, the general random effects models lead to an emphasis being placed on accurately estimating the autocorrelation function at the expense of the marginal distributions and fixed effects parameters. Therefore poor NSE scores and QQ plots can result. The QQ plots for the boreholes with strong autocorrelation were improved when the number of calibration data was doubled. The number of observations required to accurately estimate all parameters of the MM will depend on the complexity of the model and the amount of autocorrelation amongst the residuals. Therefore, it is not possible to give general guidelines about the data requirements and fitted MMs should be carefully validated to confirm their adequacy.

The QQ plots using independent validation data tended to be further from the 1:1 line than those based on the calibration data. This could be due in part to changes in the accuracy of the rainfall data over time. Jackson et al. (2016) discuss how the density of UK rain gauges varies over time.

There are two more substantive implications of assuming independent and Gaussian random effects when estimating random effects. First, the method will fail to identify temporal autocorrelation amongst the random effects. Significant temporal correlation was identified for all four boreholes and for two of the boreholes this continued for ranges up to 20 months. If this temporal correlation is not modelled then it will not be accounted for when using observations to condition predictions of GWLs (Eqn. 12). For example, if predictions of the GWL were required one month prior to the observational record, one would expect them to be correlated to the first few observations from the record and predictions which ignored the autocorrelation would be suboptimal. It is also important to account for temporal autocorrelation when simulating GWLs on multiple dates. If realisations of the hydrograph are produced where the monthly GWLs are erroneously independent, the uncertainty of the duration of events such as droughts that span multiple months will be poorly estimated. The second implication of inappropriate assumptions in the random effects is that the parameters of the fixed effects will be poorly estimated.

5.3 Formal and informal approaches to quantifying uncertainty of groundwater levels

The formal likelihood methods applied here considered the effects of parameter uncertainty and the combined effects of model specification, input and measurement errors that are included in the random effects. If the random effects model assumptions are appropriate then these are calculated using objective statistical methods. In contrast the informal uncertainty methods as implemented by Jackson et al. (2016) and Mackay et al. (2014) associate all of the predictive uncertainty with the parameter uncertainty. A subjective threshold is placed on the NSE or a similar criterion to decide whether a proposed model is behavioural. The plots of NSEs realized from the MCMC analysis of our models (Figure 11) suggest that a single NSE threshold is unlikely to be suitable for all boreholes. Indeed, the ensembles of behavioural parameters identified by Mackay et al. (2014) suggest they are considerably less identifiable than those in Figure 9. However, we note that despite these misgivings, the containment ratios recorded by Jackson et al. (2016) are comparable to those that can be inferred from our QQ plots, that the informal methodologies can be implemented more quickly than our formal likelihood approaches and that no assumptions about the structure of the model errors are required.

5.4 Further generalisations of the random effects models

Although the random effects models applied in this paper are substantially more flexible than standard independent Gaussian models there are many ways in which they could be further generalised. For example, Schoups & Vrugt (2010) permitted the variability of runoff models to linearly increase according to the flow. The variance of the random effects might also be permitted to vary according to GWLs, seasonally or according to any relevant covariate (Marchant et al., 2009). Such changes can easily be incorporated into the MM (Eqn. 1). It is possible to incorporate any marginal distribution for the random effects into the copula-based framework (Eqn. 3). The fit of the MM might also be improved by incorporating a non-Gaussian dependence structure into this framework. Eltahir and Yeh (1999) noted that groundwater drought episodes tend to last longer than flood episodes. This suggests that the correlation between successive random effects during droughts might be stronger than that during floods. Such behaviour could be accommodated by using a non-Gaussian and non-symmetric copula model (Bárdossy and Li, 2008) for the dependence structure. Before applying any of these generalisations it will be necessary to confirm that they lead to a substantial improvement in the likelihood and AIC and to confirm by validation that the resultant predictive distributions are appropriate.

6. Conclusions

Mixed models estimated by formal likelihood methods can be used to predict GWLs and to estimate the uncertainty of these predictions. In contrast to informal methods, the criterion used to estimate the models is objective and based on the likelihood that the observed data would have been realized from the specified model. However, these likelihoods are only appropriate if the assumptions on which they are based are appropriate. Therefore it is necessary to thoroughly validate the estimated MM through methods such as predictive QQ plots which assess the accuracy of the entire predictive distribution rather than just the accuracy of the estimated expected GWL at each time. If the validation results are poor then generalisations of the random effects model should be considered. GWLs recorded a month apart can be highly correlated and therefore a substantial number of observations may be required to accurately estimate all of the components of the MM. Our tests of the MM on four UK GWL hydrographs indicated that the assumptions of independent and Gaussian errors are unlikely to be completely appropriate. However, the application of these inappropriate models did not lead to a substantial deterioration of the GWL predictions or the estimates of their uncertainty. The appropriateness of the random effects model is more important in circumstances where the temporal correlation of the random effects or the posterior distributions of the fixed effects parameters are of interest.

Acknowledgements

The work described has been funded by the British Geological Survey (Natural Environment Research Council) and the Natural Environment Research Council funded project ‘Analysis of historic drought and water scarcity in the UK: a systems-based study of drivers, impacts and their interactions’ (NE/L010151//). This paper is published with the permission of the Executive Director of the British Geological Survey (Natural Environment Research Council).

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: B.N. Petov and F. Csaki (eds) Second International Symposium on Information Theory, Akademia Kiado, Budapest, 267—281.

Bárdossy, A., Li, J., 2008. Geostatistical interpolation using copulas. Water Resources Research, 44, W07412.

Beven, K., Binley, A., 1992. The future of distributed models – model calibration and uncertainty prediction. Hydrological Processes, 6, 279—298.

Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. Journal of Hydrology, 249, 1—29.

Beven, K., Smith, P.J., Freer, J.E., 2008. So just why would a modeler choose to be incoherent. Journal of Hydrology, 354, 15—32.

Bloomfield J.P., Marchant B.P., 2013. Analysis of groundwater drought using a variant of the Standardised Precipitation Index. Hydrology and Earth Systems Science, 17, 4769—4787.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. Journal of the Royal Statistical Society. Series B (Methodological), 26, 211—252.

Chandler, R., Scott, M., 2011. Statistical Methods for Trend Detection and Analysis in the Environmental Sciences. Statistics in Practice, Wiley, UK.

Daliakopoulos, I.N., Coulibaly, P., Tsanis, I.K., 2004. Groundwater level forecasting using artificial neural networks. Journal of Hydrology, 309, 229—240.

Diggle, P.J., Ribeiro, P.J., 2007. Model-based Geostatistics. Springer, New York.

Dobson, A.J., 1990. An Introduction to Generalized Linear Models. 2nd Edition. Chapman and Hall, UK.

Eltahir, E.A.B., Yeh, P. J-F., 1999. On the asymmetric response of aquifer water level to floods and droughts in Illinois, Water Resources Research, 35, 1199-1217.

Environment Agency. 2014. National groundwater level database for England, Environment Agency. http://data.gov.uk/data last retrieved 26 August 2014.

Field, M., 1983. The meteorological office rainfall and evaporation calculation system – MORECS. Agricultural Water Management, 6, 297—306.

Hastings, W.K., 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika 57, 97–109.

Jackson, C.R., Wang, L., Pachocka, M., Mackay, J.D., Bloomfield J.P., 2016. Reconstruction of multi-decadal groundwater level time series using a lumped conceptual model. Hydrological Processes. DOI:10.1002/hyp.10850.

Jackson, C.R., Pachocka, M., Mackay, J., 2013. Hydrological Outlook: Outlook based on modelled groundwater levels and seasonal climate forecast. British Geological Survey Open Report OR/13/046, Nottingham, UK (http://www.hydoutuk.net/files/2013/9566/7747/HO_methodology_Groundwater_Levels.pdf, last accessed March 2016).

Jackson, C.R., Bloomfield, J.P., Mackay, J.D., 2015. Evidence for changes in historic and future groundwater levels in the UK. Progress in Physical Geography, 39, 49—67.

Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environmental Modelling and Software, 21, 602-614.

Kazianka, H., Pilz, J., 2010. Copula-based geostatistical modelling of continuous and discrete data including covariates. Stochastic Environmental Research and Risk Assessment, 661-673.

Keller, V., Young, A.R., Morris, D., Davies, H., 2005. Continuous Estimation of River Flows (CERF), Technical Report: Estimation of Precipitation Inputs, Environment Agency R&D Project Report WD-101, Centre for Ecology and Hydrology, Wallingford.

Kidmose, J., Refsgaard, J.C., Troldborg, L., Seaby, L.P., Escrivà, M., 2013. Climate change impact on groundwater levels: ensemble modelling of extreme values. Hydrology and Earth Systems Sciences, 17, 1619—1634.

Kuczera, G., 1983. Improved parameter inference in catchment models: 1.Evaluating parameter uncertainty, Water Resources Research, 19, 1151—1162.

Lessels, J.S., Bishop, T.F.A., 2013. Estimating water quality using linear mixed models with stream discharge and turbidity. Journal of Hydrology, 498, 13—22.

Mackay, J.D., Jackson, C.R., Wang, L., 2014. A lumped conceptual model to simulate groundwater level time-series. Environmental Modelling and Software, 61, 229—245.

Marchant, B.P., Lark, R.M. 2007. The Matérn variogram model: Implications for uncertainty propagation and sampling in geostatistical surveys. Geoderma, 140, 337–345.

Marchant, B.P., S. Newman, S., Corstanje, R., Reddy, K.R., Osborne, T.Z., Lark, R.M., 2009. Spatial monitoring of a non-stationary soil property: Phosphorus in a Florida water conservation area. European Journal of Soil Science, 60(5), 757-769.

Marchant, B.P., Saby, N.P.A., Jolivet, C.C., Arrouays, D., Lark, R.M., 2011. Spatial prediction of soil properties with copulas. Geoderma, 162, 327-334.

Marshall, L., Nott, D., Sharma, A., 2005. Hydrological model selection: A Bayesian alternative. Water Resources Research, 45, W04418.

Mathworks, 2014. MATLAB and Statistics Toolbox Release 2014b, The MathWorks, Inc., Natick, Massachusetts, United States.

Minasny, B., Vrugt, J.A., McBratney, A.B., 2011. Confronting uncertainty in model-based geostatistics using Markov Chain Monte Carlo simulation. Geoderma, 163, 150—162.

Minasny, B., Whelan, B.M., Triantafilis, J., McBratney, A.B., 2013. Pedometrics research in the vadose zone-Review and perspectives. Vadose Zone Journal, 12, 4.

Mirzavand, M., Ghazavi, R., 2014. A stochastic modelling technique for groundwater level forecasting in an arid environment using time series methods. Water Resources Management, 29, 1315—1328.

Monteith, J.L., Unsworth, M.H., 2008. Principles of Environmental Physics: Third Edition. Elsevier, London, UK.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I – A discussion of principles. Journal of Hydrology, 10, 282—290.

National Groundwater Level Archive. http://www.ceh.ac.uk/data/nrfa/data/ngla.html, last access: March 2015.

Peterson, T.J., Western, A.W., 2014. Nonlinear time-series modelling of unconfined groundwater head. Water Resources Research, 50, 8330—8355.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S., 2010. Understanding predictive uncertainty in hydrologic modelling: The challenges of identifying input and structural errors. Water Resources Research, 46, W05521.

Schoups, G., Vrugt, J.A., 2010. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. Water Resources Research, 46, W10531.

Shepley, M.G., Soley, R.W.N., 2012. The use of groundwater levels and numerical models for the management of a layered, moderate-diffusivity aquifer. Geological Society, London, Special Publications, 364, 303—318.

Sun, Y., Kang, S., Li, F., Zhang, L., 2009. Comparison of interpolation methods for depth to groundwater and its temporal and spatial variations in the Minqin oasis of northwest China. Environmental Modelling and Software, 24, 1163—1170.

Stein, M.L., 1999. Interpolation of Spatial Data: Some Theory for Kriging. Springer-Verlag, New York

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S., Srikanthan, S., 2009. Critical evaluation of parameter consistency and predictive modelling in hydrological modelling: A case study using Bayesian total error analysis. Water Resources Research, 45, W00B14.

von Asmuth, J.R., Bierkens, M.F., 2005. Modelling irregularly spaced residual series as a continuous stochastic process. Water Resources Research, 41, W12404.

Vrugt, J.A., ter Braak, C.J.F., Clark, M.P., Hyman, J.M., Robinson, B.A., 2008. Treatment of input uncertainty in hydrologic modeling: doing hydrology backward with Markov chain Monte Carlo simulation. Water Resources Research 44, W00B09.

Vrugt, J.A., ter Braak, C.J.F, Diks, C.G.H, Robinson, B.A., Hyman, J.M., Higdon, D., 2009. Accelerating Markov Chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. International Journal of Nonlinear Science and Numerical Simulation, 10, 273-290.

Vrugt, J.A., 2016. Markov Chain Monte Carlo simulation using DREAM software package: Theory, concepts, and Matlab implementation. Environmental Modelling and Software, 75, 273-316.

Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists, 2nd ed., John Wiley and Sons, Chichester, UK.

Zhu, D., Zinde-Walsh, V., 2009. Properties and estimation of asymmetric exponential power distribution. Journal of Econometrics, 148, 86-99.

Figures

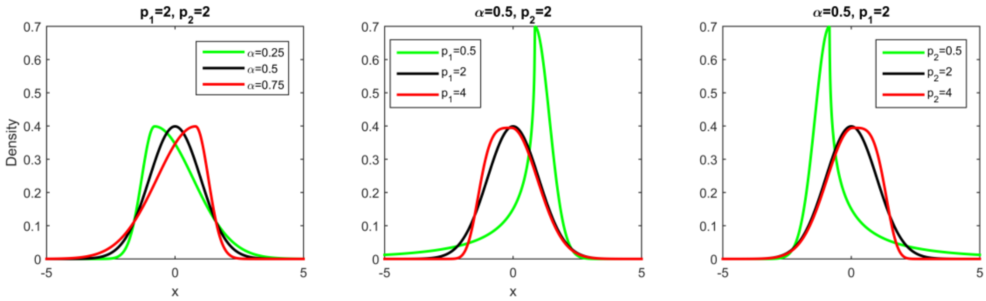


Figure 1: Examples of the AEP pdf for $\mu = 0$, $\sigma = 1$ and stated values of α, p_1 and p_2.

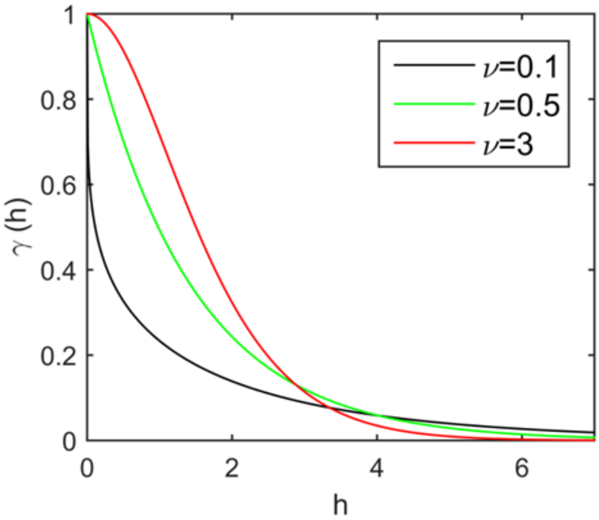


Figure 2: Examples of the Matérn covariance function with $c_1 = 1$, $\rho = 2$ and stated values of ν.

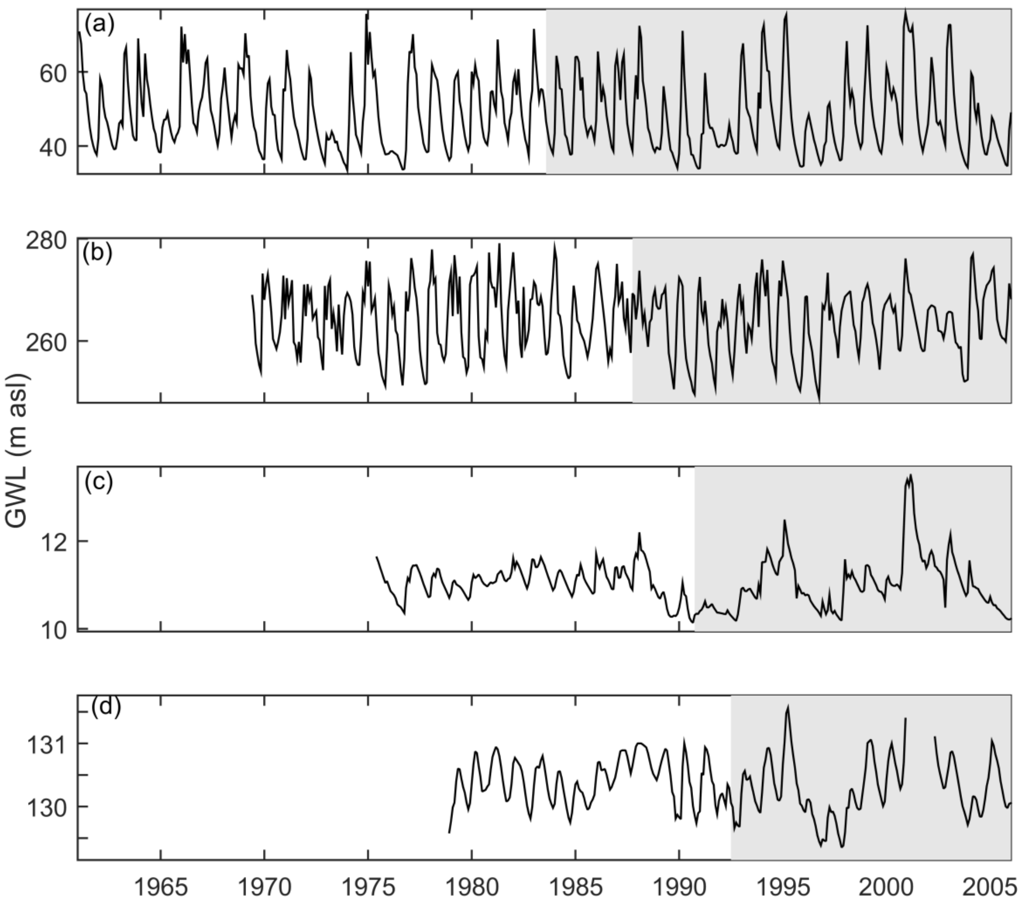


Figure 3: Observed monthly GWLs from (a) Chilgrove House, (b) Hucklow South, (c) Lower Barn Cottage, (d) Skirwith. Validation period is shaded grey.

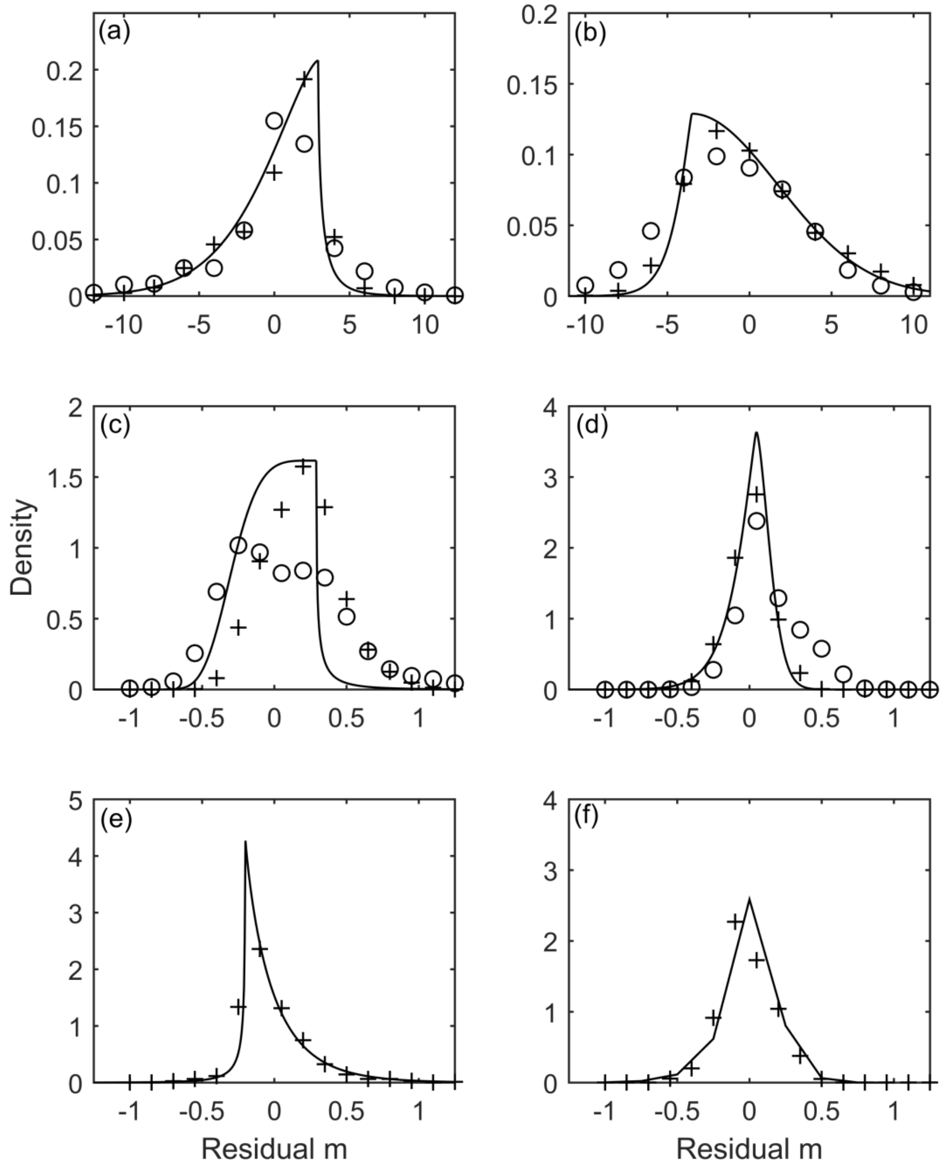


Figure 4: Maximum likelihood estimate of AEP pdf of residuals (continuous curve) and observed distribution of residuals during calibration (crosses) and validation (circles) periods. Boreholes are (a) Chilgrove House, (b) Hucklow South, (c) Lower Barn Cottage, (d) Skirwith. Plots (e) and (f) correspond to models at Lower Barn Cottage and Skirwith that have been calibrated on all of the available data.

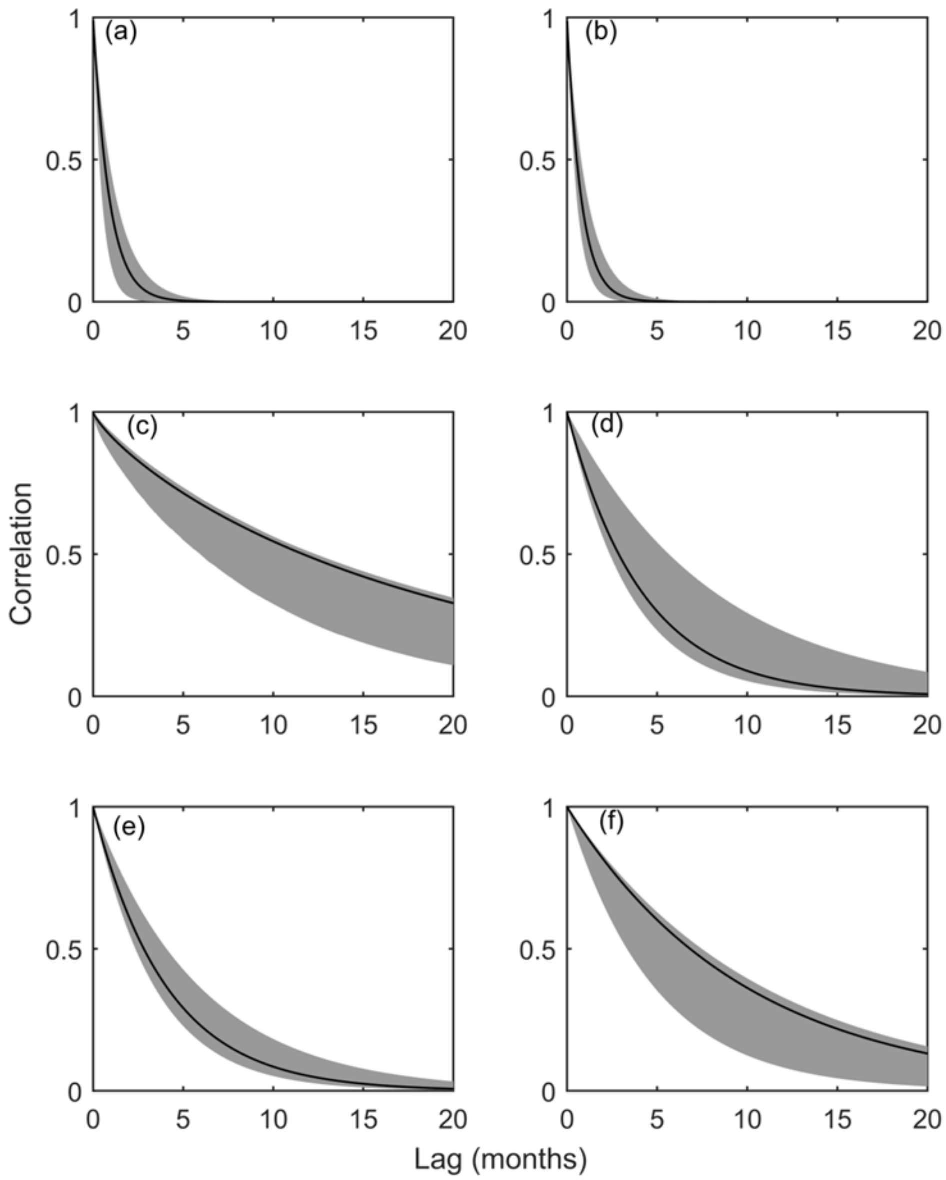
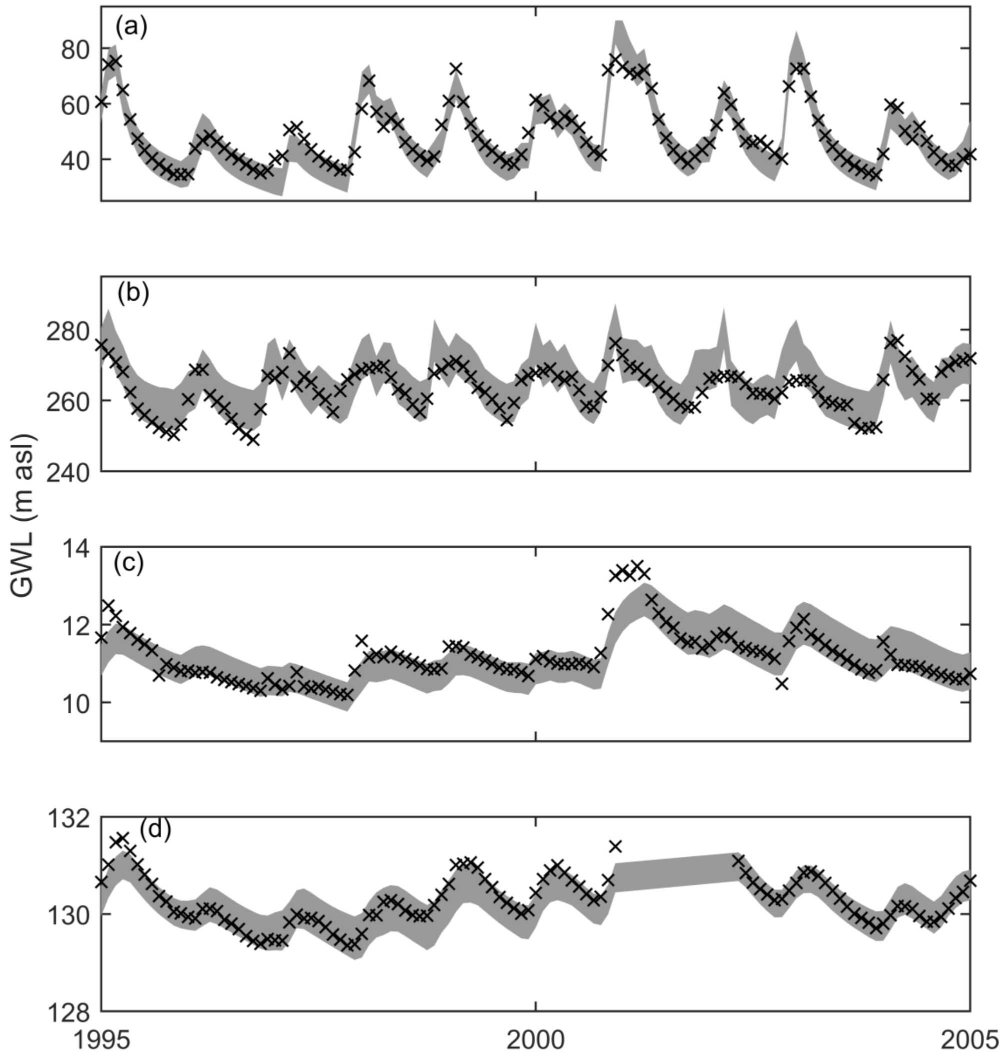


Figure 5: Maximum likelihood estimate of auto-correlation function for residuals (black line) and 90% confidence interval of the correlation function according to the MCMC sample (shaded region). Boreholes are (a) Chilgrove House, (b) Hucklow South, (c) Lower Barn Cottage, (d) Skirwith. Plots (e) and (f) correspond to models at Lower Barn Cottage and Skirwith that have been calibrated on all of the available data.



.

Figure 6: 90% prediction intervals of GWLs during a 10-year part of the validation period (shaded region) and observed GWLs (crosses) according to the best fitting MM. Prediction intervals are based on the MCMC samples and do not use conditioning data. Boreholes are (a) Chilgrove House, (b) Hucklow South, (c) Lower Barn Cottage, (d) Skirwith.

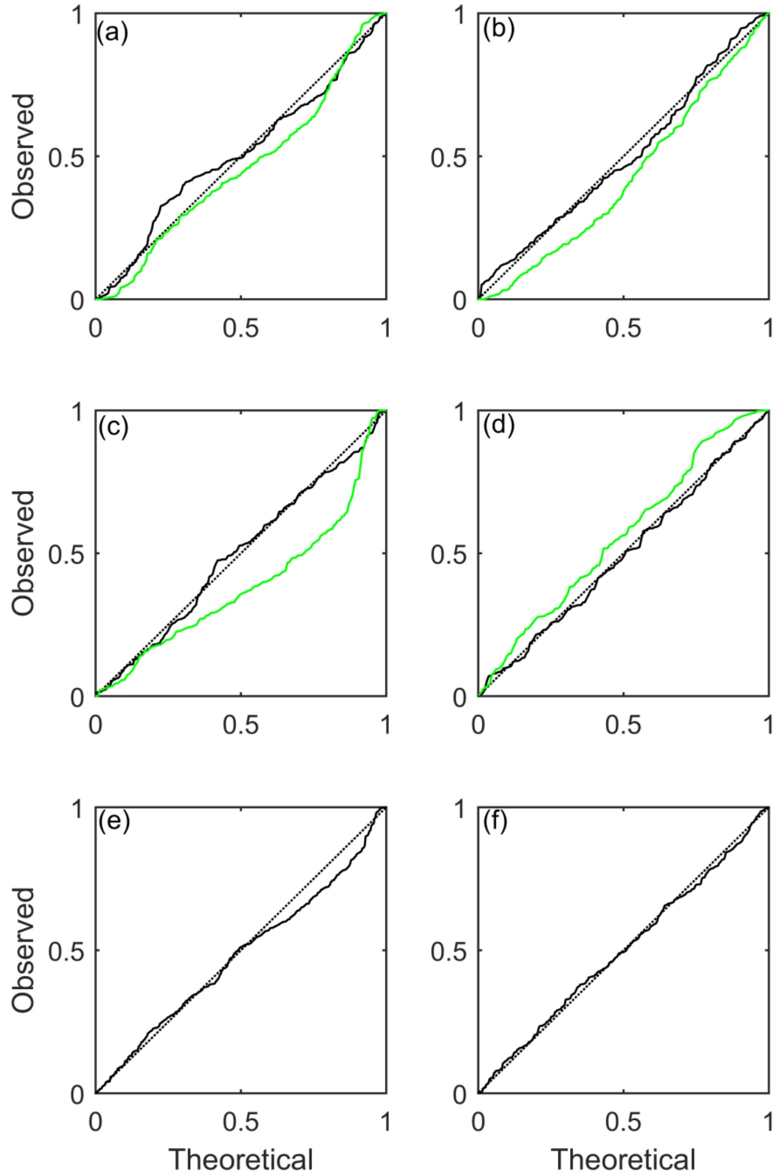


Figure 7: QQ plots upon prediction of GWLs using maximum likelihood estimate of mixed model with Gaussian and independent random effects during calibration period (black line) and validation period (green line). Boreholes are (a) Chilgrove House, (b) Hucklow South, (c) Lower Barn Cottage, (d) Skirwith. Plots (e) and (f) correspond to models at Lower Barn Cottage and Skirwith that have been calibrated on all of the available data.

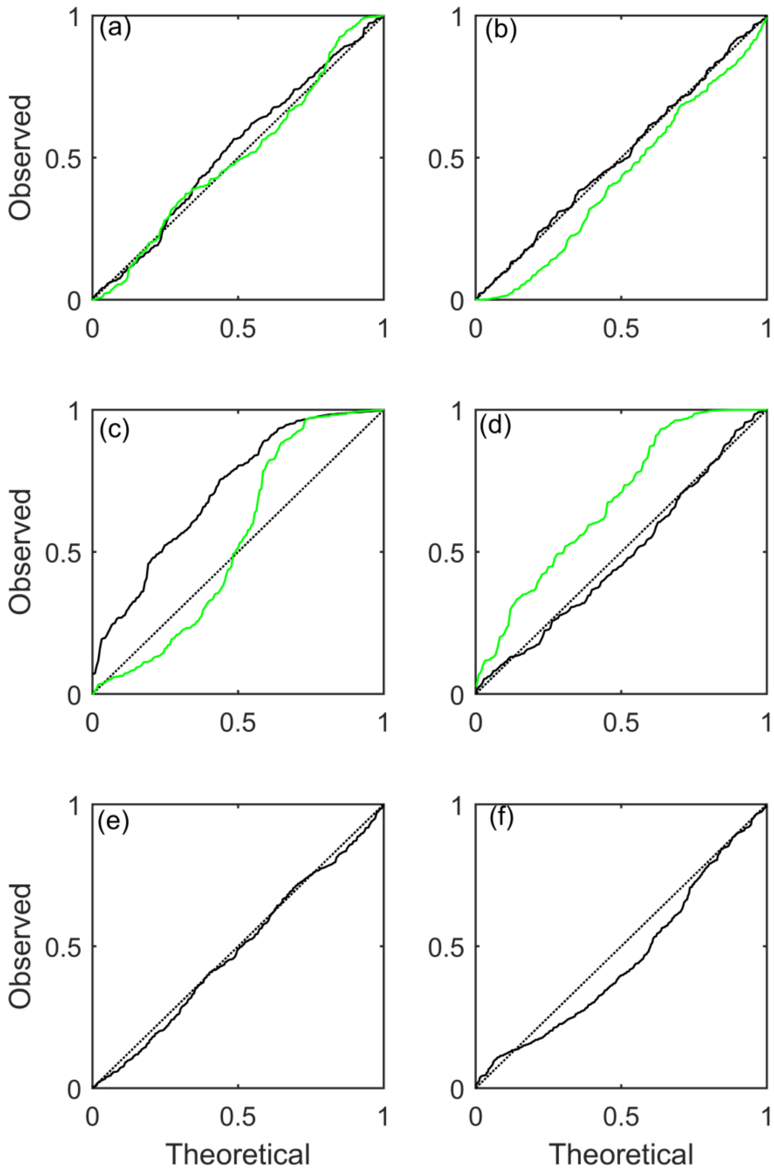


Figure 8: QQ plots upon prediction of GWLs using maximum likelihood estimate of best fitting generalized mixed model during calibration period (black line) and validation period (green line). Boreholes are (a) Chilgrove House, (b) Hucklow South, (c) Lower Barn Cottage, (d) Skirwith. Plots (e) and (f) correspond to models at Lower Barn Cottage and Skirwith that have been calibrated on all of the available data.

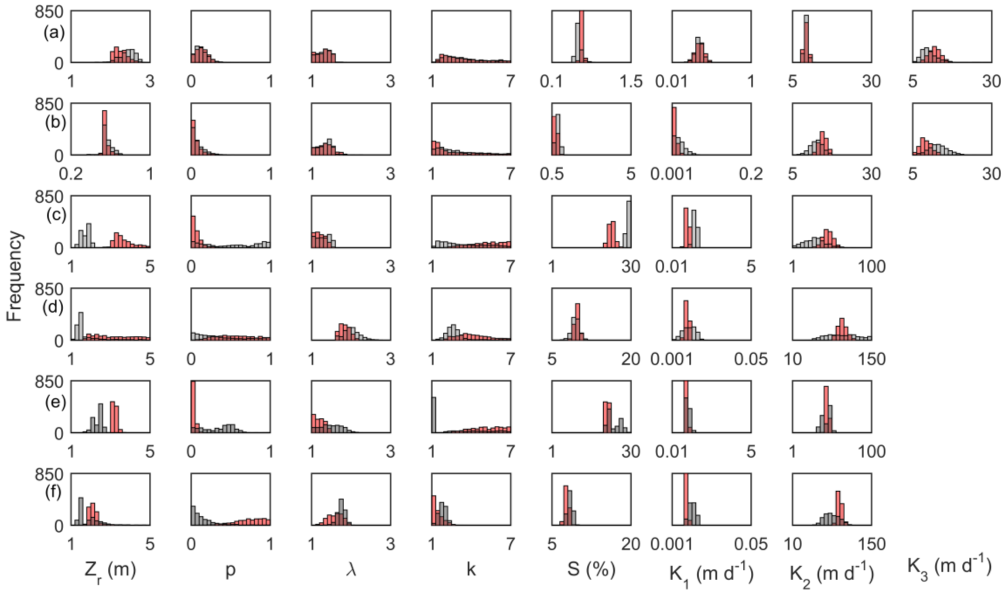


Figure 9: Histograms of AquiMod parameters realized within the 1000 MCMC samples for models with Gaussian independent (red) and AEP generalized (grey) random effects. The bounds on the parameter values correspond to the bounds on the uniform prior distributions. The boreholes are (a) Chilgrove House, (b) Hucklow South, (c) Lower Barn Cottage, (d) Skirwith. Plots (e) and (f) correspond to models at Lower Barn Cottage and Skirwith that have been calibrated on all of the available data.

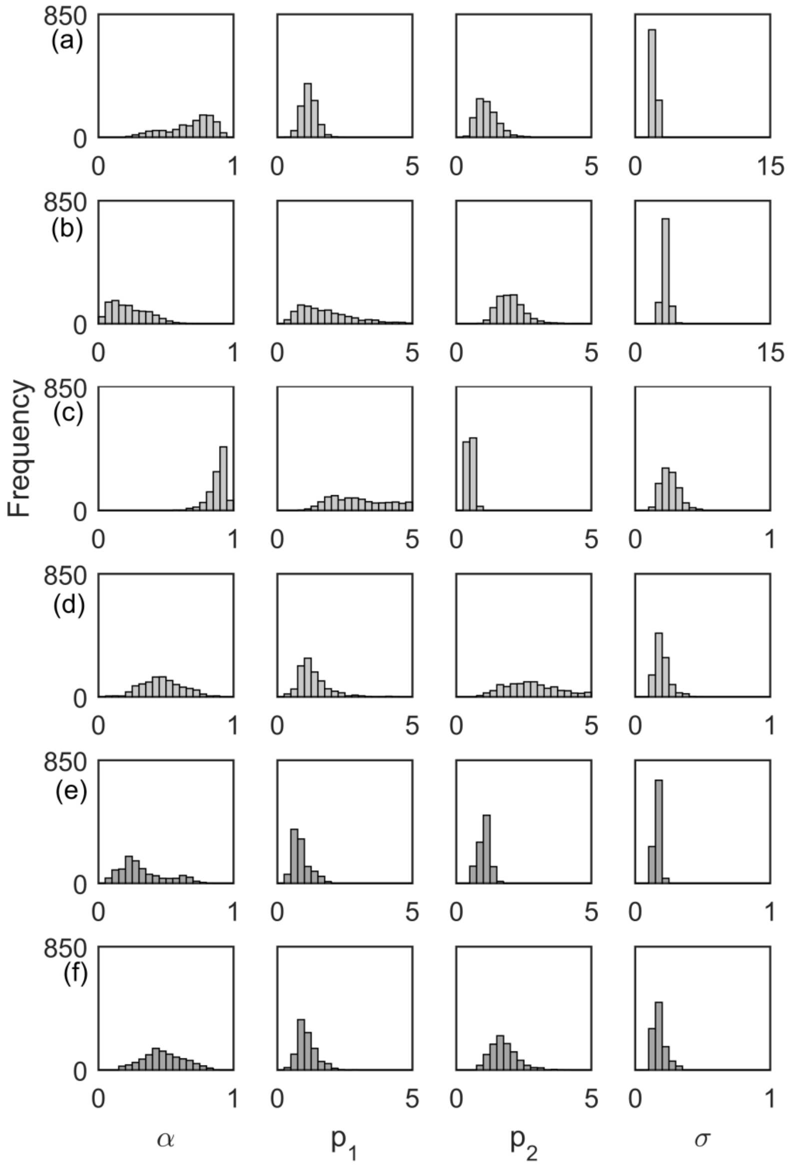


Figure 10: Histograms of AEP marginal distribution parameters realized within the 1000 MCMC samples. The bounds on the parameter values correspond to the bounds on the uniform prior distributions. The boreholes are (a) Chilgrove House, (b) Hucklow South, (c) Lower Barn Cottage, (d) Skirwith. Plots (e) and (f) correspond to models at Lower Barn Cottage and Skirwith that have been calibrated on all of the available data.

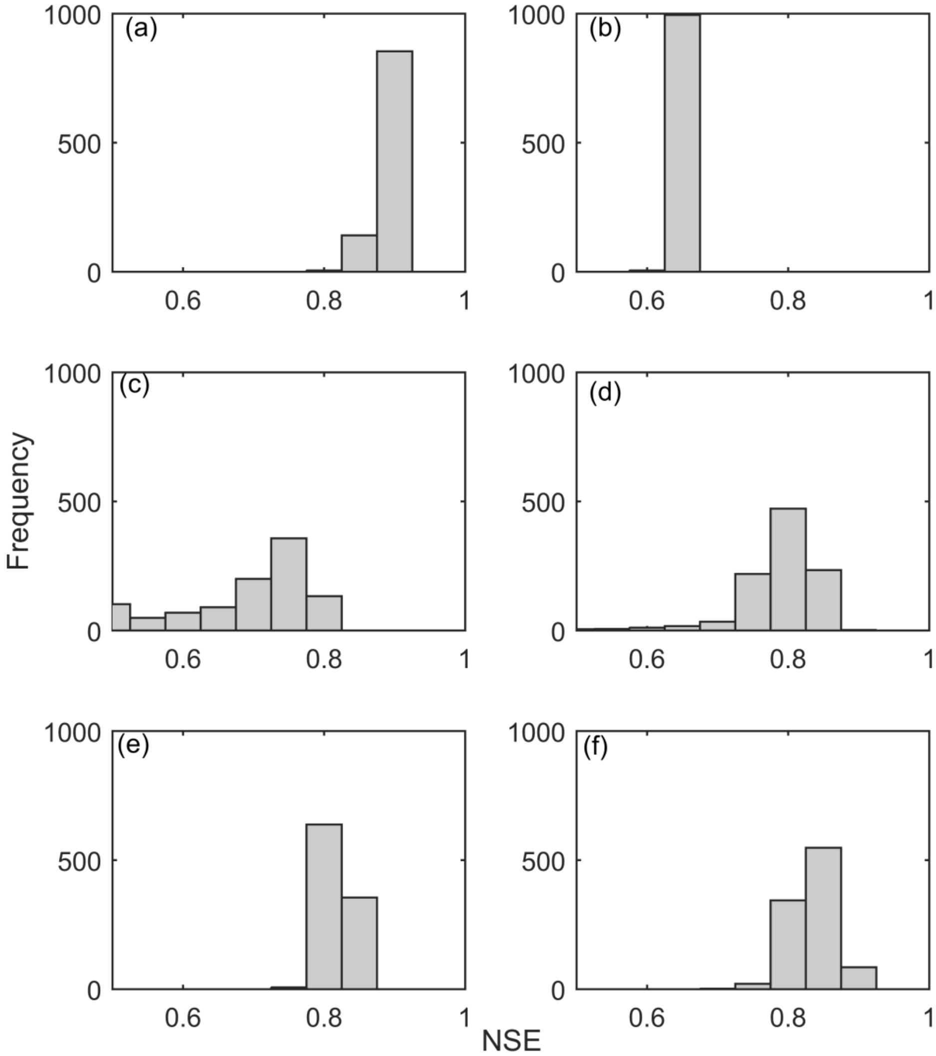


Figure 11: Calibration NSE scores for the MCMC samples of AquiMod parameters. The boreholes are (a) Chilgrove House, (b) Hucklow South, (c) Lower Barn Cottage, (d) Skirwith. Plots (e) and (f) correspond to models at Lower Barn Cottage and Skirwith that have been calibrated on all of the available data.

	Gaussian			AEP		
	Independent	Exponential	Matérn	Independent	Exponential	Matérn
Chilgrove House	1247.2	1241.5	1243.1	1219.3	**1209.1**	1211.1
Hucklow South	1137.1	1133.3	1135.5	1122.5	**1113.9**	1117.1
Lower Barn Cottage	-87.1	-237.0	-243.8	-108.4	-294.1	**-294.6**
Skirwith	-162.1	-277.7	-275.4	-155.6	**-283.9**	-280.8

Table 1: AIC values for maximum likelihood estimates of mixed models for the calibration data from four boreholes with different distributions and correlation functions. Smallest AIC values are highlighted in bold.

	Gaussian			AEP		
	Independent	Exponential	Matérn	Independent	Exponential	Matérn
Chilgrove House	0.93	0.93	0.93	0.93	0.93	0.90
Hucklow South	0.74	0.73	0.73	0.73	0.73	0.70
Lower Barn Cottage	0.74	0.65	0.64	0.76	0.76	0.54
Skirwith	0.84	0.82	0.83	0.84	0.84	0.78

Table 2: NSE scores for maximum likelihood estimates of mixed models for the calibration data from the four boreholes with different distributions and correlation functions.